RESEARCH ARTICLE SUMMARY

NONHUMAN GENOMICS

Long-read sequence assembly of the gorilla genome

David Gordon,* John Huddleston,* Mark J. P. Chaisson,* Christopher M. Hill,* Zev N. Kronenberg,* Katherine M. Munson, Maika Malig, Archana Raja, Ian Fiddes, LaDeana W. Hillier, Christopher Dunn, Carl Baker, Joel Armstrong, Mark Diekhans, Benedict Paten, Jay Shendure, Richard K. Wilson, David Haussler, Chen-Shan Chin, Evan E. Eichler†

INTRODUCTION: The accurate sequence and assembly of genomes is critical to our understanding of evolution and genetic variation. Despite advances in short-read sequencing technology that have decreased cost and increased throughput, whole-genome assembly of mammalian genomes remains problematic because of the presence of repetitive DNA.

RATIONALE: The goal of this study was to sequence and assemble the genome of the western lowland gorilla by using primarily single-molecule, real-time (SMRT) sequencing technology and a novel assembly algorithm that takes advantage of long (>10 kbp) sequence reads. We specifically compare the properties of this assembly to gorilla genome assemblies that were generated by using more routine short sequence read approaches in order to determine the value and biological impact of a long-read genome assembly.

RESULTS: We generated 74.8-fold SMRT whole-genome shotgun sequence from peripheral blood DNA isolated from a western lowland gorilla (Gorilla gorilla gorilla) named Susie. We applied a string graph assembly algorithm, Falcon, and consensus algorithm, Quiver, to generate a 3.1-Gbp assembly with a contig N50 of 9.6 Mbp. Short-read sequence data from an additional six gorilla genomes was mapped so as to reduce indel errors and improve the accuracy of the final assembly. We estimate that 98.9% of the gorilla euchromatin has been assembled into 1854 sequence contigs. The assembly represents an improvement in contiguity: >800-fold with respect to the published gorilla genome assembly and >180-fold with respect to a more recently released upgrade of the gorilla assembly. Most of the sequence gaps are now closed, considerably increasing the yield of complete gene models. We estimate that 87% of the missing exons and 94% of the incomplete genes are recovered. We find that the sequence of most full-length common repeats is resolved, with the most significant gains occurring for the longest and most G+C-rich retrotransposons. Although complex regions such as the major histocompatibility locus are accurately sequenced and assembled, both heterochromatin and large, high-identity segmental duplications are not because read lengths are insufficiently long to traverse these repetitive structures. The long-read assembly produces a much finer map of structural variation down to 50 bp in length, facilitating the discovery of thousands of lineage-specific structural variant differences that have occurred since divergence from the human and chimpanzee lineages. This includes the disruption of specific genes and loss of predicted regulatory regions between the two species. We show that use of the new gorilla genome assembly changes estimates of divergence and diversity, resulting in subtle but substantial effects on previous population genetic inferences, such as the timing of species bottlenecks and changes in the effective population size over the course of evolution.

CONCLUSION: The genome assembly that results from using the long-read data provides a more complete picture of gene content, structural variation, and repeat biology, improving population genetic and evolutionary inferences. Long-read sequencing technology now makes it practical for individual laboratories to generate high-quality reference genomes for complex mammalian genomes.

The list of author affiliations is available in the full article online.
*These authors contributed equally to this work.
†Corresponding author. E-mail: eee@gs.washington.edu

Cite this article as D. Gordon et al., Science 352, aae0344 (2016). 10.1126/science.aee0344

A Susie, reference sample

B Long-read assembly (Susie3)

C Short-read assembly (gorGor3)

Long-read sequence assembly of the gorilla genome. (A) Susie, a female western lowland gorilla, was used as the reference sample for full-genome sequencing and assembly [photograph courtesy of Max Block]. (B and C) Treemaps representing the differences in fragmentation of the long-read and short-read gorilla genome assemblies. The rectangles are the largest contigs that cumulatively make up 300 Mbp (~10%) of the assembly.
Long-read sequence assembly of the gorilla genome

David Gordon,1,2* John Huddleston,1,2* Mark J. P. Chaisson,1,2 Christopher M. Hill,1,2 Zev N. Kronenberg,1,2 Katherine M. Munson,1 Maika Mallig,1 Archana Raja,1,2 Ian Fiddes,2 LaDeana W. Hillier,4 Christopher Dunn,3 Carl Baker,1 Joel Armstrong,3 Mark Diekhans,3 Benedict Paten,3 Jay Shendure,1,2 Richard K. Wilson,4 David Haussler,3 Chen-Shan Chin,5 Evan E. Eichler1,2†

Accurate sequence and assembly of genomes is a critical first step for studies of genetic variation. We generated a high-quality assembly of the gorilla genome using single-molecule, real-time sequence technology and a string graph de novo assembly algorithm. The new assembly improves contiguity by two to three orders of magnitude with respect to previously released assemblies, recovering 87% of missing reference exons and incomplete gene models. Although regions of large, high-identity segmental duplications remain largely unresolved, this comprehensive assembly provides new biological insight into genetic diversity, structural variation, gene loss, and representation of repeat structures within the gorilla genome. The approach provides a path forward for the routine assembly of mammalian genomes at a level approaching that of the current quality of the human genome.

High-quality sequence and assembly of genomes is a Lynchpin to our understanding of the genetic diversity and evolution of species. The development of massively parallel sequencing technologies has drastically reduced the cost and increased the throughput of genome sequencing (1, 2). Although these advances have enabled sequencing of many more species and individuals, assemblies are left incomplete and fragmented in large part because the underlying sequence reads are too short (≤200 base pairs [bp]) to traverse complex repeat structures (3). This has led to incomplete gene models, less accurate representation of repeats, and biases in our understanding of genome biology. The published western lowland gorilla genome assembly, for example, was generated by a mix of ABI capillary sequence and whole-genome shotgun Illumina short sequencing read pairs (4). Although the accuracy is high and many important inferences regarding the evolution of the species could be made, the resulting assembly contains more than 400,000 sequence gaps (Table 1) (4). Moreover, the use of the human genome to help guide the assembly of the gorilla genome created an artificially low number of structural rearrangements (5). The effect of these misassemblies and missing data becomes exacerbated when genomic comparisons are made among multiple primate genomes. For example, nearly 20% of the human genome could not be readily aligned in a four-way comparison among apes in large part because of the draft nature of nonhuman primate genomes (2). The recent development and application of long-read sequencing technologies has shown considerable promise in improving human genome assemblies as well as our understanding of genetic variation (3, 6, 7). We applied this long-read sequencing technology to generate an alternate genome assembly of the western lowland gorilla.

Table 1. Gorilla assembly statistics.

<table>
<thead>
<tr>
<th></th>
<th>gorGor3*</th>
<th>Susie3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual</td>
<td>Kamilah</td>
<td>Susie</td>
</tr>
<tr>
<td>Total genome length (bp)</td>
<td>3,035,660,144</td>
<td>3,080,414,926</td>
</tr>
<tr>
<td>Number of contigs</td>
<td>464,874</td>
<td>16,073</td>
</tr>
<tr>
<td>Total sequence (bp)</td>
<td>2,828,866,575</td>
<td>3,080,414,926</td>
</tr>
<tr>
<td>Placed contig length (bp)</td>
<td>2,718,960,062</td>
<td>2,790,620,487</td>
</tr>
<tr>
<td>Unplaced contig length (bp)</td>
<td>109,906,513</td>
<td>289,794,439</td>
</tr>
<tr>
<td>Maximum contig length (bp)</td>
<td>191,556</td>
<td>36,219,563</td>
</tr>
<tr>
<td>Contig N50 (bp)</td>
<td>11,661</td>
<td>9,558,608</td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>57,196</td>
<td>554</td>
</tr>
<tr>
<td>Maximum scaffold length</td>
<td>10,247,101*</td>
<td>110,018,866</td>
</tr>
<tr>
<td>Scaffold N50 (bp)</td>
<td>913,458</td>
<td>23,141,960</td>
</tr>
</tbody>
</table>

*Values are taken from previously published gorilla genome paper (4).

†Corresponding author. E-mail: eee@gs.washington.edu
published gorilla genome reference (gorGor3), we closed 94% of the 433,861 gorGor3 gaps (fig. S13), resulting in the addition of at least 164 Mbp of euchromatic sequence. This additional sequence dramatically improved gene annotation, including the discovery of thousands of exons and putative regulatory elements (Fig. 1C).

An analysis of the gaps in gorGor3 showed that they were enriched 3.8-fold for Alu short interspersed nuclear element (SINE) repeats, revealing that this G+C-rich primate repeat was particularly problematic in the initial gorilla assembly. Overall, there was a positive correlation between gap size and repeat content, especially for segmental duplications (three- to fivefold enrichment). For example, 10,959 gorGor3 gaps were in excess of 2 kbp; of these, 21% (2298) mapped to segmental duplications (table S12). Although heterochromatic regions could still not be resolved in the new gorilla assembly, our analysis of the underlying sequence data indicates that 10% of the gorilla genome consists of satellite repeats (table S15). A 32-bp satellite (pCht7) (9) associated with subterminal caps of gorilla chromosomes was more abundant than α-satellite repeats (4.0 versus 2.3%) within the genome sequence data (table S15). In contrast to other ape species, these data suggest that telomeric-associated heterochromatic repeats are more abundant than centromeric repeats in gorillas.

Quality assessment
We assessed the quality of the assembly using two independent sources of data. First, we used paired-end sequence data generated during the sequencing and assembly of the first western lowland gorilla (Kamilah) to assess the assembly integrity. Our analysis showed that 98.6% of the new gorilla assembly was supported by concordant paired-end sequence data generated from either large-insert BAC or fosmid gorilla clones (table S2). An analysis of aligned high-quality Sanger sequence data (54.5 Mbp) revealed high sequence identity (99.71%) within the potential range of allelic diversity expected among western lowland gorillas (10) (table S10). In order to more precisely assess sequencing accuracy, we generated and aligned Illumina whole-genome sequence data from the same gorilla (Susie) and used sequence differences to estimate an initial error of one per 1000 base pairs [quality value (QV) = 30]. The most frequent errors observed were a threefold excess of single base pair insertions. We used Illumina whole-genome sequence data from Susie as well as six additional western lowland gorilla genomes (10) to correct these indel errors. Using these additional genomes, we created a “pan” gorilla reference genome (Susie3), correcting indel errors and identifying the most common single-nucleotide polymorphism variants. After error correction, we estimate that Susie3 has less than one error per 5000 bp (QV > 35) (table S7).

Comparative analyses between human and gorilla genomes
To determine the potential biological utility of this new reference gorilla genome, we measured the increased yield in gene content from gap closures in the gorilla reference by identifying human RefSeq (17) exons that map within gap regions of gorGor3 that are closed in Susie3. Using this projection, we estimate that 87% (11,105 of 12,754) of the missing RefSeq exons are recovered and that 94% of the incomplete gorilla genes are resolved for at least one isoform in Susie3. More generally, we found that an additional 1 to 3% of illumina RNA-sequencing (RNA-seq) data and an additional 6 to 7% of gorilla ESTs are mapped by standard mapping methods to this more complete genome assembly (table S13). The assembly also improves the contiguous representation of genes predicted by human GENCODE (12). Compared with other nonhuman primate genomes, Susie3 has far fewer assembly errors, making it one of the most complete primate genomes after human (Fig. 4). We found that 88% of protein-coding transcripts (GENCODE) have >99% identity against Susie3 compared with ~55% in gorGor3 (fig. S18). We leveraged the new assembly, available RNA-seq data, and Augustus TransMap (13–15) alignments to define a new set of 45,087 consensus gene models for protein-coding transcripts (isoforms from 19,633 genes) in the gorilla genome.

The greater sequence contiguity of the genome assembly provided an opportunity to comprehensively assess structural variation between gorilla and human genomes at a fine-scale resolution, down to ~50 bp. Comparison of the gorilla (Susie3) and human (GRCh38) assemblies revealed a total of 117,512 insertions and deletions (92 Mbp) and 697 inversion variants (Table 2 and table S20). More than 86% of these events were previously unidentifed (fig. S47) (6, 17), and 72% of the events were determined to represent fixed differences specific to the gorilla lineage. Variants ranged in size from small structural variants within the coding sequence of genes to large, complex, gene-rich structural variant events spanning hundreds of kilobase pairs (Fig. 4, C

![Fig. 1. Gorilla genome assembly.](http://science.sciencemag.org)
such gene-disrupting events may represent polymorphisms or be enriched for errors in a reference, we restricted our analysis to previously undescribed events that were shared across an additional six western lowland gorillas (table S19).

Using these criteria, we identified 146 structural variants (76 deletions and 69 insertions) that affect the coding sequence of 110 distinct RefSeq genes. Although no single functional gene category reached statistical significance for enrichment, several of the genetic differences between the species were associated with sensory perception, keratin production, interleukin and cytokine secretion, reproduction, immunity, growth, transmembrane signaling, and nucleotide binding (table S19).

We also investigated structural variants that affected potential regulatory elements on the basis of recent annotations of the human genome (ENCODER). We annotated base pair–resolved variants that intersected with deoxyribonuclease I hypersensitive (DNase I HS) sites associated with open chromatin, histone H3 lysine 4 trimethylation (H3K4me3) marks associated with transcriptionally active regions/promoters, and histone H3 acetylation on lysine 27 (H3K27ac) signals associated with enhancers (22). Among the 10,466 insertions and deletions that intersected regulatory elements, 2151 represented fixed structural differences between humans and gorillas (GSVs).

Similarly, we identified 133 GSVs that affect long intergenic noncoding RNA (lincRNA). As expected, we observed a significant (2- to 10-fold) depletion of structural variants intersecting these functional categories, with the exception of lincRNA.

We further quantified the spatial correlation between GSVs and the regulatory elements by comparing their proximity. The midpoints of the GSVs overlapped significantly less with the protein-coding genes (projection test \( P = 6.0 \times 10^{-12} \)) and with fetal DNase I HS sites (projection test \( P = 4.3 \times 10^{-35} \)), which is consistent with the action of purifying selection. We observed a modest spatial enrichment between GSVs that map within 50 to 200 bp of putative promoter and enhancer marks (projection test \( P = 0.041 \) and \( 4.1 \times 10^{-15} \) for H3K4me3 and H3K27ac, respectively).

This analysis identified 327 protein-coding genes within 10 kb of an H3K4me3 mark that overlaps or maps near a GSV (<100 bp) and 672 genes that had the same spatial pattern for H3K27ac marks. We compared these fixed structural differences within putative regulatory DNA with 127 genes with differences in RNA expression and CTCF binding sites between human and gorilla (20). We identified nine genes in which a GSV intersected an H3K4me3 or H3K27ac mark (ADAMTS10, ALDH1L1, CDH1, COL5A1, GRK5, IGFBP1, INSR, IQGAP2, and SRC) and eight that intersected the 3’ untranslated region or a noncoding transcribed exon (AMOTL1, DUSP4, HNF4B, ITGB5, SGC22, TCL1A, ZDHHC19, and ZNF907). None of these GVSs affected protein-coding regions of these genes, and thus, these fixed differences in noncoding regulatory DNA are strong candidates to explain the expression differences between the species.

Six inversions were identified that potentially disrupt an exon or break within a gene, in addition to 206 insertions and microinversions that map within introns. Among the insertion and deletion events, 15 of the structural variants occurred in genes that are largely intolerant to mutation in humans (residual variation intolerance score < 20 percentile) (22). This includes insertions within the insulin-degrading enzyme (IDE) and a negative regulator of WNT signaling (AXIN1), as well as deletions in a tubulin-specific chaperone (TBCD) and an extracellular sulfatase (SULF2) associated with cell signaling. Our analysis revealed a large number of in-frame deletions and insertions that result in the addition and loss of amino acids and protein domain differences between human and gorilla genes. Although the impact of these structural variants awaits further characterization, the greater resolution of structural variants provides a large number

---

**Fig. 2. Gorilla genome ideogram.** Schematic depicting assembly contig lengths mapped to gorilla chromosomes. The first two rows of black rectangles represent contigs >3 Mbp, the green rectangles correspond to contigs >1 Mbp and ≤3 Mbp, and blue rectangles correspond to contigs ≤1 Mbp.
Table 2. Gorilla genome structural variants. Contigs greater than 200 kbp were mapped to GRCh38 by using BLASR (30). Nonrepetitive sequences contained at most 70% of sequence annotated as repeat by RepeatMasker (3.3.0) (28) or Tandem Repeats Finder 4.07b (29). Mosaic repeats are defined as one or more different repeat annotations; however, mosaic structural variants composed solely of Alu are listed separately because of their frequency.

<table>
<thead>
<tr>
<th>Repeat element</th>
<th>Count</th>
<th>Fixed</th>
<th>Average length (bp)</th>
<th>Standard deviation (bp)</th>
<th>Total bases</th>
<th>Count</th>
<th>Fixed</th>
<th>Average length (bp)</th>
<th>Standard deviation (bp)</th>
<th>Total bases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complex, not repetitive</td>
<td>15,383</td>
<td>13,190</td>
<td>709.31</td>
<td>1807.29</td>
<td>1091284</td>
<td>13,749</td>
<td>10,250</td>
<td>763.17</td>
<td>1,954</td>
<td>10492765</td>
</tr>
<tr>
<td>Complex, repetitive</td>
<td>7,952</td>
<td>7,243</td>
<td>2716.96</td>
<td>2746.63</td>
<td>20518484</td>
<td>7,450</td>
<td>5,934</td>
<td>2757.39</td>
<td>2906.61</td>
<td>20542529</td>
</tr>
<tr>
<td>AluY</td>
<td>6,418</td>
<td>5,878</td>
<td>284.36</td>
<td>67.2</td>
<td>1825012</td>
<td>9,422</td>
<td>7,882</td>
<td>294.58</td>
<td>60.04</td>
<td>2775515</td>
</tr>
<tr>
<td>AluS</td>
<td>990</td>
<td>950</td>
<td>160.64</td>
<td>85.98</td>
<td>159036</td>
<td>1,053</td>
<td>845</td>
<td>159.84</td>
<td>82.05</td>
<td>168315</td>
</tr>
<tr>
<td>L1Hs</td>
<td>86</td>
<td>50</td>
<td>262.67</td>
<td>582.37</td>
<td>22590</td>
<td>220</td>
<td>179</td>
<td>235.51</td>
<td>358.93</td>
<td>51813</td>
</tr>
<tr>
<td>CR1</td>
<td>1,339</td>
<td>1,256</td>
<td>438.5</td>
<td>737.1</td>
<td>587147</td>
<td>1,181</td>
<td>874</td>
<td>450.84</td>
<td>744.1</td>
<td>532441</td>
</tr>
<tr>
<td>LIP</td>
<td>1,891</td>
<td>1,660</td>
<td>278.26</td>
<td>487.86</td>
<td>526192</td>
<td>1,430</td>
<td>1,098</td>
<td>312.81</td>
<td>493.88</td>
<td>447317</td>
</tr>
<tr>
<td>SVA</td>
<td>1,498</td>
<td>1,097</td>
<td>1561.2</td>
<td>1007.55</td>
<td>2338672</td>
<td>1,194</td>
<td>786</td>
<td>1279.32</td>
<td>914.18</td>
<td>1527511</td>
</tr>
<tr>
<td>HERV</td>
<td>324</td>
<td>285</td>
<td>241.06</td>
<td>498.09</td>
<td>78104</td>
<td>260</td>
<td>194</td>
<td>362.79</td>
<td>797.94</td>
<td>94325</td>
</tr>
<tr>
<td>PTEVR</td>
<td>190</td>
<td>179</td>
<td>2973.22</td>
<td>3427</td>
<td>564912</td>
<td>1</td>
<td>1</td>
<td>188</td>
<td>0</td>
<td>188</td>
</tr>
<tr>
<td>Mosaic-Alu</td>
<td>1,433</td>
<td>1,355</td>
<td>320.11</td>
<td>262.59</td>
<td>458717</td>
<td>1,514</td>
<td>1,177</td>
<td>307.92</td>
<td>260.53</td>
<td>466194</td>
</tr>
<tr>
<td>STR</td>
<td>11,049</td>
<td>4,520</td>
<td>367.74</td>
<td>813.01</td>
<td>2703261</td>
<td>7,828</td>
<td>4,225</td>
<td>323.38</td>
<td>558.5</td>
<td>2531424</td>
</tr>
<tr>
<td>Satellite</td>
<td>7,351</td>
<td>4,520</td>
<td>367.74</td>
<td>813.01</td>
<td>2703261</td>
<td>7,828</td>
<td>4,225</td>
<td>323.38</td>
<td>558.5</td>
<td>2531424</td>
</tr>
<tr>
<td>Other</td>
<td>1,319</td>
<td>1,184</td>
<td>201.01</td>
<td>247.8</td>
<td>265130</td>
<td>1,196</td>
<td>896</td>
<td>180.49</td>
<td>227.79</td>
<td>215872</td>
</tr>
<tr>
<td>Not base-pair resolved</td>
<td>41</td>
<td>N/A</td>
<td>49.496</td>
<td>2381.46</td>
<td>2029323</td>
<td>61</td>
<td>N/A</td>
<td>55.971</td>
<td>43.496</td>
<td>3414209</td>
</tr>
<tr>
<td>Total</td>
<td>58,621</td>
<td>44,653</td>
<td>789.54</td>
<td>2183.43</td>
<td>46283738</td>
<td>58,891</td>
<td>40,482</td>
<td>775.55</td>
<td>2804.21</td>
<td>45672675</td>
</tr>
</tbody>
</table>

Fig. 3. Comparison of gorilla genome assemblies. The contig length distribution for the resulting long-read assembly (Susie3) is 2 to 3 orders of magnitude larger when compared with previous gorilla genome assemblies (gorGor3 and gorGor4) that were generated by using Illumina and Sanger sequencing technology.

Sequence and assembly of the major histocompatibility locus
As a test of the enhanced structural variation detection and its potential biomedical relevance, we compared the organization of the major histocompatibility complex (MHC) II locus between gorilla and human (GRCh37). MHC encodes genes that are critical for antigen presentation on immune cells, and its content and structure are known to differ radically between closely related primates (22). We found that segmental duplication has expanded the locus in gorilla relative to human so that in the assembly, there are 79,166 (~10%) duplicated base pairs, compared with 53,084 bp (~8%) in human. We identified three large gorilla insertions across the ~1 Mbp region (15, 48, and 52 kbp) that correspond to 14% of the locus (Fig. 4C), including several previously unidentified MHC genes. In terms of assembly statistics, the MHC region in gorGor3 has 168 gaps that are filled in Susie3, eight of which mapped within 100 bp of 27 distinct genes (human RefSeq).

To verify the organization of the HLA region in Susie3, we sequenced a tiling path of BACs derived from another gorilla, Kamilah, who previously contributed to the published gorGor3 assembly (table S6). We generated two high-quality, clone-based sequence haplotypes for the region (MHC haplotype 1, 863,324 bp; MHC haplotype 2, 289,560 bp). Both Kamilah MHC haplotypes mapped with higher sequence identity to Susie3 (99.6 and 99.9%) than gorGor3 (98.1 and 98.3%), which is consistent with assembly improvements for this complex region of the genome.

Population genetic inferences based on the new assembly
Last, we assessed whether the new gorilla genome assembly would have any influence on previous population genetic inferences as a result of potential changes in estimates of divergence and diversity. Although the difference was subtle, we found that human versus gorilla sequence alignments were significantly less divergent (P < 2.2 × 10^-16; Welch two-sample t test) with Susie3 (1.60% divergent) when compared with the published gorilla assembly (1.65% divergent) (Fig. 6). Both regional and chromosomal analyses showed that the different estimates of gorilla-human divergence were nonrandomly distributed. In particular, we observed that specific gene-rich regions and chromosomes (such as chromosomes 19 and
of single-nucleotide variants for six western lowland gorillas against Susie3 and gorGor3. Although a greater fraction of the Illumina paired-end reads mapped with higher mapping quality to Susie3 (mean, 0.75) compared with Susie3 (mean, 0.33). Next, we examined the average observed heterozygosity across both assemblies in 100-kbp windows (fig. S8). The largest difference in observed heterozygosity between the assemblies was on the X chromosome, in which the average gorGor3 heterozygosity was 0.27 versus 0.23 in Susie3.

It is likely that the increased gorGor3 heterozygosity is due to mapping errors owing to fewer reads successfully mapping (gorGor3 has more than 400,000 gaps, many of which contain repetitive sequences). Mapping software errors on the side of sensitivity, meaning reads derived from underrepresented regions may be incorrectly placed. To test this idea, we identified the reads (including mate-pairs) from Coco, a female gorilla, corresponding to heterozygous positions in gorGor3 and remapped them to both gorGor3 and Susie3. We found that only 87% of the heterozygous calls corresponding to heterozygous positions in gorGor3 resulted in heterozygous calls in Susie3.

We also assessed the effect of the new assembly on estimates of diversity by comparing the pattern of single-nucleotide variants for six western lowland gorillas against Susie3 and gorGor3. Although a greater fraction of the Illumina paired-end reads mapped with higher mapping quality to Susie3 (table S8), we observed a higher ratio of heterozygous genotypes in gorGor3 (mean, 0.35) as compared with Susie3 (mean, 0.33). Next, we examined the average observed heterozygosity across both assemblies in 100-kbp windows (fig. S62). The largest difference in observed heterozygosity between the assemblies was on the X chromosome, in which the average gorGor3 heterozygosity was 0.27 versus 0.23 in Susie3.

It is likely that the increased gorGor3 heterozygosity is due to mapping errors owing to fewer reads successfully mapping (gorGor3 has more than 400,000 gaps, many of which contain repetitive sequences). Mapping software errors on the side of sensitivity, meaning reads derived from underrepresented regions may be incorrectly placed. To test this idea, we identified the reads (including mate-pairs) from Coco, a female gorilla, corresponding to heterozygous positions in gorGor3 and remapped them to both gorGor3 and Susie3. We found that only 87% of the heterozygous calls in gorGor3 resulted in heterozygous calls in Susie3 (table S33). The “lost” heterozygous genotypes correspond to regions with lower Illumina read depth as compared with that of gorGor3, supporting the notion that mapping error is likely inflating heterozygosity in the original assembly (fig. S63).

Accurate estimates of population histories are important for understanding how climate change, disease, and human activity shaped the genetic diversity within western lowland gorillas. Such estimates are critically dependent on genetic measures such as heterozygosity. We fit a pairwise sequentially Markovian coalescent (PSMC) model to Illumina data from four western lowland gorillas mapped against Susie3 and gorGor3 (fig. 6). Comparing the two assemblies, we observed two statistically significant differences in the predicted effective population size of the gorilla species at ~50 thousand years ago (ka) and ~5 million years ago (ma) (Fig. 6). The effects occur in opposite directions. We observed a significant decrease in the effective population size at more ancient time periods (P < 0.0001; Welch two-sample t test), possibly because of a greater fraction of repetitive sequence and segmental duplications being resolved in Susie3. In contrast, use of Susie3 as a reference predicts a significantly greater effective population size at 50 ka (P < 0.0001; Welch two-sample t test). Our results suggest that the most recent bottleneck of the western lowland gorilla
Fig. 5. Improved mobile element resolution. (Left) PTERV1 and SVA insertion length and percent identity distributions in Susie3 (blue) and gorGor3 (red). The PTERV1 and SVA elements in gorGor3 are biased toward short but on average higher identity alignments to the consensus sequence because the more divergent long terminal repeat sequences are not resolved. (Right) The mean and median insertion lengths for gorGor3 and Susie3 are PTERV1, 2194.93, 7565.85 (median 1223 and 7725) and SVA, 1240.1, and 1965.63 (median 1162 and 1909).

Fig. 6. Population genetic analyses. (A) Density of average divergence within 1-Mbp windows between human (GRCh38) and gorGor3, Susie3, or chimpanzee (panTro4) autosomes. (B) A comparison of human-gorGor3 and human-Susie3 divergence over 1-Mbp windows. The x axis is Alu coverage in each window, and the y axis is the difference in human-gorilla divergence between gorGor3 and Susie3. Positive y axis values indicate increased human–Susie3 divergence relative to human–gorGor3. The increased divergence of human–gorGor3 correlates with Alu content (slope, −0.0044094; intercept, 0.0001486; Pearson’s correlation, −0.60). (C) The effective population size (N_e) shown over time. A PSMC model was applied to the western lowland gorilla based on different genome assemblies. Illumina genome sequence data from western lowland gorillas (Abe, Amani, Coco, Tzambo) was mapped against gorGor3 (green) and Susie3 (orange), and PSMC was fit to the genome alignments (-N25 -t15 -r5 -b -p "4+25*2+4+6"; mutation rate = 1.25 × 10^{-8}; generation time = 19 years). There are 100 bootstrap replicates for each gorilla and model. (D) The distribution of the bootstrap intervals that overlap 50 ka and 5 ma. At 50 ka, Susie3 estimates of the effective population size are significantly higher than that for gorGor3; the inverse pattern is true for 5 ma. All differences between Susie3 and gorGor3 are significant (**∗∗∗P ≤ 0.0001; Welch two-sample t test).
aligning Illumina (HiSeq 2500 and NextSeq) sequence data generated from six additional western lowland gorillas (90), including 14-fold sequence data generated from the new gorilla reference genome (Susie). Consensus gene models were built with Augustus (15) by use of TransMap (14, 15) alignments of GENCODE transcripts to Susie3 and previously published RNA-seq data (26, 27).

REFERENCES AND NOTES
Long-read sequence assembly of the gorilla genome

David Gordon, John Huddleston, Mark J. P. Chaisson, Christopher M. Hill, Zev N. Kronenberg, Katherine M. Munson, Maika Malig, Archana Raja, Ian Fiddes, LaDeana W. Hillier, Christopher Dunn, Carl Baker, Joel Armstrong, Mark Diekhans, Benedict Paten, Jay Shendure, Richard K. Wilson, David Haussler, Chen-Shan Chin and Evan E. Eichler

Science 352 (6281), aae0344.
DOI: 10.1126/science.aee0344

Improving on the gorilla genome
Access to complete, high-quality genomes of nonhuman primates will also help us understand human biology. Gordon et al. used long-read sequencing technology to improve genome data on our close relative the gorilla. Sequencing from a single individual decreased assembly fragmentation and recovered previously missed genes and noncoding loci. Mapping short-read sequences from additional gorillas helped reconstruct a “pan” gorilla sequence documenting genetic variation. Comparison with human genomes revealed species-specific differences ranging in size from one to thousands of bases in length, including some that are likely to affect gene regulation.

Science, this issue p. 10.1126/science.aee0344

ARTICLE TOOLS
http://science.sciencemag.org/content/352/6281/aae0344

SUPPLEMENTARY MATERIALS
http://science.sciencemag.org/content/suppl/2016/03/30/352.6281.aae0344.DC1

REFERENCES
This article cites 62 articles, 15 of which you can access for free
http://science.sciencemag.org/content/352/6281/aae0344#BIBL

PERMISSIONS
http://www.sciencemag.org/help/reprints-and-permissions